Arellano-Bond Model

Xiang Ao

October 30, 2007

1 Dynamic Panel Data Model

A static panel data model takes the form

$$y_{it} = \mathbf{x}_{it}\beta_1 + \nu_i + \epsilon_{it}, \tag{1}$$

where ν_i can be random or fixed effect for unit i.

A dynamic panel data model allows past realizations of the dependent variable to affect its current level. Consider the model

$$y_{it} = \sum_{j=1}^{p} \alpha_j y_{i,t-j} + \mathbf{x}_{it}\beta_1 + \mathbf{w}_{it}\beta_2 + \nu_i + \epsilon_{it}, \tag{2}$$

where ν_i can be random or fixed effect for unit i.

Two new terms are introducted: $\mathbf{w}_{it}$ is a vector of pre-determined covariates, which can be lags of x_{it}'s. $\sum_{j=1}^{p} \alpha_j y_{i,t-j}$ are lagged dependent variables, up to p lags.

$\mathbf{w}_{it}$ does not have a big impact on how to estimate the model. Given the condition that they are predetermined, a regular panel data estimator (for example, a fixed effect estimator) is still consistent. However, the introduction of $\sum_{j=1}^{p} \alpha_j y_{i,t-j}$ makes a regular panel data estimator inconsistent. To see why it is so, let's consider a fixed effect estimator (within estimator). Suppose we have

$$y_{it} = \alpha_j y_{i,t-1} + \mathbf{x}_{it}\beta_1 + \mathbf{w}_{it}\beta_2 + \nu_i + \epsilon_{it}, \tag{3}$$

which means, we only have $y_{i,t-1}$ on the right hand side.

A within estimator regresses $(y_{it} - \bar{y}_i)$ on $(y_{i,t-1} - \bar{y}_i)$ and other "demeaned" regressors, with an erorr term $(\epsilon_{it} - \bar{\epsilon}_i)$. The within estimator gets rid of ν_i. However, $y_{i,t}$ correlates with ϵ_{it}, so $y_{i,t-1}$ correlates with $\epsilon_{i,t-1}$. Therefore $y_{i,t-1}$ correlates with $\bar{y}_i$. So $(y_{i,t-1} - \bar{y}_i)$ correlates with $(\epsilon_{it} - \bar{\epsilon}_i)$, which makes the within estimator inconsistent. This inconsistency goes away if $\bar{\epsilon}_i$ is very small comparing to ϵ_{it}, which can occur in a long panel.

Let's consider a simpler situation, with $p = 2$; that is, only two lags of y_i are included in the regressors. Also let's omit $\mathbf{w}_{it}$ for now. Consider first-differencing equation 2,

$$y_{it} - y_{i,t-1} = \alpha(y_{i,t-1} - y_{i,t-2}) + (\mathbf{x}_{it} - \mathbf{x}_{i,t-1})'\beta + (\epsilon_{it} - \epsilon_{i,t-1}), \qquad (4)$$

As we discussed, OLS estimator is inconsistent, since $y_{i,t-1}$ is correlated with $\epsilon_{i,t-1}$.

Anderson and Hsiao (1981) proposed using instrumental variables estimator with $y_{i,t-2}$ as an instrument for $(y_{i,t-1} - y_{i,t-2})$.

Arellano and Bond (1991) expanded the idea by using additional lags of the dependent variable as instruments. For example, both $y_{i,t-2}$ and $y_{i,t-3}$ can be uased as instruments. In fact, as t increases, the number of instruments avialable also increases. In period 3 only $y_{i,1}$ is available. In period 4 $y_{i,1}$ and $y_{i,2}$ are available. In period 5 $y_{i,1}$ and $y_{i,2}$ and $y_{i,3}$ are availabe, and so on. In other words, we'll have an instrument matrix with one row for each time period that we are instrumenting:

$$Z_i = \begin{bmatrix} y_{i,1} & y_{i,2} & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \Delta x_{i,4} \\ 0 & 0 & y_{i,1} & y_{i,2} & y_{i,3} & \cdots & 0 & 0 & 0 & \Delta x_{i,5} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \cdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 0 & y_{i,1} & \cdots & y_{i,T-2} & \Delta x_{i,T} \end{bmatrix} \qquad (5)$$

In the case of $p = 2$, the number of rows of Z_i is $T - 3$. The number of columns of Z_i depends on the length of T, it can grow very quickly with T. In general cases, Z_i has $T - p - 1$ rows and $\sum_{m=p}^{T-2}(m + k_1)$ columns, where k_1 is the number of variables in x.

When there are endogenous variables in $\mathbf{x}$, they are treated similarly to the lagged dependent variables, and levels lagged two or more periods are valid instruments. For predetermined variables, levels lagged one or more periods are valid instruments.

Let H_i be the $(T_i - p - 1) \times (T_i - p - 1)$ covariance matrix of the differenced idiosyncratic errors:

$$H_i = E[\epsilon_i^* \epsilon_i^{*\prime}] = \begin{bmatrix} 2 & -1 & 0 & \cdots & 0 & 0 \\ -1 & 2 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 2 & -1 \\ 0 & 0 & 0 & \cdots & -1 & 2 \end{bmatrix} \qquad (6)$$

The Arellano-Bond estimator is to use the moment conditions:

$$E[Z_i' \epsilon_i^*] = 0 \text{ for } i = 1, 2, \ldots, N \qquad (7)$$

The one-step Arellano-Bond estimator is a GMM estimator using H_i:

$$\hat{\delta}_1 = Q_1^{-1}(\sum_{i=1}^{N} X_i^{*\prime} Z_i) A_1 (\sum_{i=1}^{N} Z_i' y_i^*) \qquad (8)$$

where

$$Q_1 = (\sum_{i=1}^{N} X_i^{*\prime} Z_i) A_1 (\sum_{i=1}^{N} Z_i' X_i^{*\prime}) \tag{9}$$

and

$$A_1 = (\sum_{i=1}^{N} Z_i{}' H_i Z_i)^{-1} \tag{10}$$

The two-step estimator is based on one-step estimator:

$$\hat{\delta}_2 = Q_2^{-1} (\sum_{i=1}^{N} X_i^{*\prime} Z_i) A_2 (\sum_{i=1}^{N} Z_i' y_i^*) \tag{11}$$

where

$$Q_2 = (\sum_{i=1}^{N} X_i^{*\prime} Z_i) A_2 (\sum_{i=1}^{N} Z_i' X_i^{*\prime}), \tag{12}$$

$$A_2 = (\sum_{i=1}^{N} Z_i{}' G_i Z_i)^{-1}, \tag{13}$$

and

$$G_i = \hat{\epsilon}_i^* \hat{\epsilon}_i^{*\prime}. \tag{14}$$

That is, the only difference between one-step and two-step estimator is the weighting matrix A_1 and A_2. A_2 uses the one-step residual from the one-step estimation. Please note that the one-step estimator is based on a homoskedastic error term. Simulation studies have suggested very modest efficiency gain from using the two-step version, even in the presence of considerable heteroskedasticity. More importantly the dependence of the two-step weight matrix on estimated parameters makes the usual asymptotic distribution approximations less relable for the two step estimator. Simulation studies have shown that the asymptotic standard errors tend to be much too small for the two-step estimator. These are the reasons that much applied work has focused on the results of one-step estimator.

The robust estimators for variance-covariance matrix have been developped. Here we'll skip that.

When $T > 3$ and the model is overidentified, a Sargan test can be used to test the overidentifying restrictions.

2 Arellano-Bond estimation implemented

Arellano and Bond have written a program in GAUSS to do dynamic panel data modeling:

`http://www.ifs.org.uk/publications.php?publication_id=3255`

Stata has implemented dynamic panel data models since version 8. Please note that between version 9 and 10, there is significant syntax change in xtabond.

Stata 10 now has a few commands that model dynamic panel data:

xtabond, xtdpd, and xtdpdsys.